基于局部图结构的链接预测模型

赵思云，黄增峰
(复旦大学 大数据学院，上海 200433)

摘　要：链接预测是基于已知的部分图数据来预测节点之间未被观测到的边或者未来可能产生的边的任务。链接预测领域目前最表现最佳的方法是，对所有目标节点对提取周围的低阶邻居小图，使用小图做图分类预测链接的方法。然而，这种方法的稳定性和性能受限于图的局部结构特异性。提出的方法在上述算法的基础上进行了改进。该算法根据目标节点周围节点的结构特征计算周围节点优先值，根据优先值筛选出高优先值的节点集合，并同时选出一定数量的随机节点，共同组成封闭子图，提取子图特征进行链接预测。实验表明，该算法有效提高了在不同结构的图数据上选出的小图的精准性和稳定性，显著提升了链接预测的效果。
关键词：链接预测；子图提取；PageRank；节点编号
中图分类号：TP391.4　　doi: 10.19734/j.issn.1001-3695.2022.03.0117

Link prediction method based on local topological structure

Zhao Siyun, Huang Zengfeng
(*School of Data Science*, *Fudan University*, *Shanghai* 200433, *China*)

Abstract: Link prediction is a task of predicting unobserved edges between nodes or edges that may be connected in the future based on partial graph data. The current state-of-art method of link prediction is to extract the surrounding low-hop subgraphs for all target node pairs and perform graph classification algorithm on the subgraphs to predict the focal link. However, its stability and performance are limited by the diversity of local topological structures. This paper proposed a method to improve the above algorithm. The algorithm calculated the priority value of the surrounding nodes according to their topological feature, selected the most important nodes among the surrounding nodes and a certain number of random nodes to form a closing subgraph together, then extracted feature from the closing subgraph to predict the link. Experiments show that the algorithm ensures the accuracy and stability of intelligently extracting subgraphs on graph data of different structures, and significantly improves the accuracy of link prediction.
Key words: link prediction; subgraph extraction; PageRank; node labeling

0　引言

在高度信息化的现代社会，数据有很多不同的表现形式，其中图数据在生物[1]、医疗[2]、社交网络[3]、知识补全[4]等领域都具有非常好的应用，而链接预测则是图数据分析中比较重要的任务之一。图数据由节点和边构成，每个节点表示不同的实体，而边则表示实体之间的各种关联。在实际情况中，图数据往往都是不完整和动态变化的，本文在某个时刻观测到的图数据可能具有片面性和时效性，所以如何依据已知的部分图数据对真实的节点关联情况进行预测就变得尤为重要。

传统的链接预测算法主要是启发式的算法，从节点的相似性出发，认为具有相似背景或者处于相似环境中的节点具有更大的倾向会建立关联关系，而在已知图中距离较远、所处拓扑环境差异较大的节点对则在直观上来看毫无联系，也就被认为建立连边的可能性更小。这一类的方法在特定的领域仍然具有很好的表现，例如，张玲玲等人[5]将启发式的算法与节点本身的特性结合，在对研发者的潜在合作者进行链接预测时取得了不错的效果。基于图嵌入学习的方法[6~8]也被用于进行链接预测任务。无监督的图嵌入算法会通过学习图中的拓扑结构，将在图上距离比较近或者关联比较紧密、邻居结构比较相似的节点赋予相近的特征向量，然后用两个节点的特征向量作为输入训练一个简单的 0-1 分类器就能比较好的对链接进行预测。在图卷积神经网络[9~11]出现之后，通过图卷积的方法，先结合邻居节点特征对每个节点的初始特征向量进行卷积变换，再用得到的新特征向量进行分类预测，将链接预测任务的效果提升了很大一个台阶。由于图卷积神经网络的卷积层数往往比较低，对于每一个节点而言，算法辐射的跳数范围比较有限，所以说明图数据的局部拓扑结构对链接预测任务具有比较高的有效性。近年来，Singh 等人[12]提出了基于边集两次预测的链接预测模型，认为原始的训练集中的边与真实数据存在较大差异的现象是影响链接预测准确性的主要原因。他们使用一种方法对训练集中的边进行一次预测补全后，再选用另一相同或不同的算法，基于补全后的边集来做链接预测。Li 等人[13]提出了基于距离增强的链接预测方法，在全图中选出一些较为重要的节点，并计算其他节点到这些节点的距离参数，将这些参数加入神经网络中进行预测。

Zhang 等人[14]提出了 SEAL 模型，该工作证明了所有启发式的算法均可用中心节点的 k 跳子图做近似，并提出了抽取目标节点对周围的邻居 k 跳小图，对小图做图分类进行链接预测的方法，也使得链接预测任务在稳健性和准确性上取得了很大的突破。

上述方法各自从不同的角度对链接预测算法进行了改善和提升，但是仍然存在一些局限性。图数据的稠密程度、全图结构特征、局部连边结构在不同背景的数据集上差异非常

收稿日期：2022-03-14；**修回日期：**2022-05-09
作者简介：赵思云(1997-)，女，江西九江人，硕士，主要研究方向为图卷积神经网络(19210980108@fudan.edu.cn)；黄增峰(1985-)，男，浙江杭州人，青年研究员，博导，主要研究方向为机器学习算法与理论、大数据计算、理论计算机科学、复杂网络分析.

大。所以本文希望，在目前表现最佳的“提取子图+图分类”的链接预测框架下，图分类端输入的子图能更加规范，这就要求它至少具有相近的节点个数。另一方面，目标节点对的周围重要程度高的节点不一定位于它们的低跳邻居里，所以本文希望更加智能的找到链接预测任务中重要程度更高的节点。

本文基于 SEAL[14]提出的链接预测框架进行了改进，提出了一种更有针对性的固定节点个数的子图提取方法，在不同稠密程度和拓扑结构的局部区域上，可以兼顾随机性和特异性的选择重要的周围节点进入封闭子图，同时相对应的调整了适合的节点编号与图分类方法，显著的提升了模型的性能。总的来说，本文的贡献主要包括以下三点：

a) 基于“提取子图+图分类”的链接预测框架，结合个性化 PageRank(personalized PageRank，PPR)等启发式方法，提出了一种端到端的链接预测模型，应对不同稠密程度和不同背景的图数据，发现周围节点对于中心节点的重要性差异，智能的对大图进行预处理和子图提取，并最终通过图分类算法得到链接预测结果。

b) 提出了一种针对目标节点对的封闭子图提取方法，综合目标节点对周围节点的全局重要性和局部重要性，使每个提取出的封闭子图具有更高的表达力和相同的规模，提高了在链接预测场景下图分类任务的输入规范性。

c) 在多个不同背景的数据集上进行大量实验，并与多个具有代表性的基线模型进行实验对比，得到了非常优秀的效果。

基于子图提取和图分类的链接预测框架如图 1 所示。

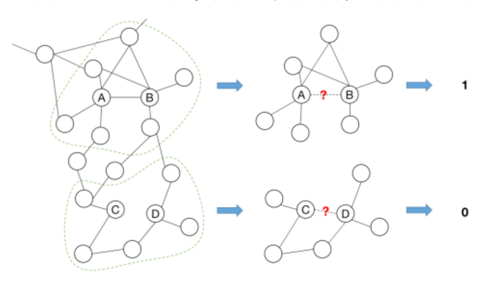


图 1　基于子图提取和图分类的链接预测框架

Fig. 1　A link prediction framework base on subgraph extraction and graph classification

1　相关工作

1.1　启发式方法

启发式的方法是最早被用来做链接预测的传统方法之一，这是基于一些可以计算的图数据上的静态特征描述节点之间的相似性，并通过这些相似性对节点间是否存在边相连进行预测的方法的统称。总的来说，这类方法认为节点相似性越高的节点对存在边的概率越高，反之越低。启发式方法可以粗略的分为一阶方法、二阶方法和高阶方法。顾名思义，一阶的启发式方法在计算过程中只需要用到两个节点之间的一阶邻居，如共同邻居个数法、Jaccard 系数法、择优连接法[15]等；二阶的启发式方法最多用到两个目标节点的二度邻居，如 AA(Adamic-Adar)[3]和 RA(resource allocation)[16]；高阶的启发式方法可以用到两个目标节点的三度及以上的所有邻居，最常见的有 PageRank[17]、SimRank[18]、Katz 系数法[19]等。启发式方法的局限性也非常明显，即基于静态图计算的指标特征在不同的数据上都有比较大的差异，而且单个的指标往往无法比较全面的衡量拓扑结构的多维特征，所以表达力度也比较有限。本文提出的模型可以基于不同数据的特点智能的训练链接预测模型，同时也综合了多个维度的启发式方法，比较全面的描述了节点之间的相关关系。

1.2　基于图嵌入方法的链接预测

图数据是一种非常高维的非欧数据结构，所以想要直接利用图网络结构中的所有信息会非常困难，而且计算代价很大。图嵌入方法[6,7,20,21]在这个时候就应运而生，它的本质是希望通过低维的向量来表达每个节点中蕴涵的图结构信息。因此，好的图嵌入方法可以在学到了图中每个节点的图嵌入向量特征之后，能够通过这些节点特征向量尽可能准确的反推出完整的图网络结构。变分图自编码器模型[7]将节点特征矩阵的每一行看做是一个高维高斯分布的随机变量，构建模型学习高斯分布的均值和方差，通过高斯分布采样得到每个节点的特征矩阵 $Z_{n\times d}$，其中 n 表示节点数量，d 表示特征维数，之后使用 $Z \cdot Z^T$ 作为解码器还原出原始的邻接矩阵。Node2vec[6]是基于随机游走的无监督图嵌入方法，它用图上的连边权重来构建从每个节点出发走到其他邻居节点的概率矩阵，然后以此在图上采样出大量随机游走序列，同时使用负采样的方式，随机抽取一些在图上相距非常远的节点对，通过优化节点特征向量的内积使得距离越近的节点特征向量越相似，而距离越远的节点特征向量越无关。因此，图嵌入方法所得到的节点特征向量往往天然与图上的连边情况息息相关，使用图嵌入方法之后，再将目标节点对的两个节点特征输入简单的分类器模型，就往往能得到很好的效果。这一类的图嵌入方法聚焦学习图网络结构，但是无法将节点的原生特征与图的拓扑结构综合到一起进行学习，所以还是损失了一定的信息和学习效率。本文提出的模型通过能够综合节点的原生特征和局部拓扑结构，很好的解决了这一问题。

1.3　图卷积神经网络

图卷积神经网络也是图数据上的一类可扩展性和表达力度都很高的模型。这一类方法的基本思想是在图结构中通过邻居关系来传递并聚合信息。一般来说，图卷积神经网络类方法会先聚合每个节点的周围所有邻居特征，再将聚合后的信息与目标节点当前的信息进行加权合并，然后使用这些信息同时更新图上所有节点的特征向量。在图卷积神经网络类的算法研究中，不同的加权方法、采样方法、聚合方法等被纳入考虑进行了研究。Kipf 等人[11]提出的 GCN 模型，通过使用均值聚合来近似计算的方式，把图的卷积操作推广到了图上的谱域上。为了解决图的动态更新问题以及不同节点邻居数量分布不均匀的问题，Hamilton 等人[9]提出了 GraphSAGE 模型，该方法采用有放回抽样的方式在每次聚合操作时对每个节点抽取相同数量的邻居节点，将所有所抽取的邻居节点特征与中心节点特征合并，并逐点更新下一层的节点特征。GAT 模型[10]在图卷积神经网络中引入了注意力机制，它考虑到聚合过程中每个邻居节点不同的相对重要性，通过学习多个注意力参数来控制聚合过程中邻居节点的相对权重，使得图卷积变得更加智能。GIN 模型[22]提出了一种新型的聚合合并方式，使得图卷积神经网络模型可以在区别同构图的问题上做到接近 Weisfeiler-Lehman 测试[23]的效果，同时也在图卷积神经网络的传统任务中达到了非常良好的性能。然而基于全图的图卷积神经网络方法由于训练时读入的视野范围非常大，而无法聚焦目标节点对周围的小图的局部拓扑结构，因此忽略了很多局部特征。本文的模型通过提取目标节点对周围的邻居小图进行训练的方式，使得模型能够更多的关注到目标节点对周围的局部网络结构的细微特征，从而更准确的对链接是否存在进行预测。

1.4　SEAL

SEAL 模型[14]是近年来最有突破性的链接预测模型之一，是目前为止在链接预测任务上表现最佳的模型，也是本文的主要对比模型之一。SEAL 开创性的提出了基于“封闭小图提取+图分类”的链接预测框架，证明了所有的高阶或低阶的启发式特征均能够用目标节点对的低阶邻居子图做近似，从而说明了对于链接预测任务而言，每一个目标节点对周围

的子图包含了进行链接预测所需要的所有高阶和低阶的特征，为“封闭小图提取+图分类”的框架提出了理论支持。同时，SEAL 提出了节点编号对于该框架的重要性，它认为邻居节点(包括直接邻居和高阶邻居)对于目标节点的重要性是各不相同的，需要在小图进行区别，因而提出了“双半径编号法”来表示在封闭小图中不同地位的节点，相同地位(编号)的节点共享同一个特征向量，这样在对封闭小图进行图分类时就可以共享相同的参数。

虽然在目标节点对周围提取封闭子图进行训练的方法在大部分数据集上都表现极佳，但是粗暴地直接取 k 跳子图的方式并不能很好的发挥出封闭子图表达力度的极限，反而可能会因为选取了无关或者比较边缘的节点，导致学习封闭子图的结构效率变低或者效果受到噪声干扰。另一方面，直接选取的 k 跳子图规模大小会随着不同目标节点对所处位置的局部连边稠密程度而改变。这也使得后续的图分类任务变得更加不规范，在全图稠密程度差异较大的情况下，选出的封闭子图中的节点数量的方差就会很大，在同一个图分类模型下的分类准确率就会进一步降低。本文提出的模型就很好的解决了这几个问题。一方面，本文提出的模型可以在不同背景、不同拓扑结构以及全图分布差异性较大的图数据上，更加智能的选出对于位于中心的目标节点对而言，重要程度排名较高的前 n 个节点。这可以使得小图的规模更加精准统一，而不是随着稠密程度和局部拓扑结构的不同而自由改变规模的大小；另一方面，本文提出的模型在提取封闭子图的过程中，使用了多个启发式的方法，在综合考量了全图信息和局部信息的同时，还保留了一定的可以调节的随机性。这使得本文的链接预测模型在收集封闭子图的时候，有能力随机地看到分布在目标节点对的周围，但是原本重要性不高的环境节点，从而保留了模型对于反常拓扑结构的一定的适应性。

2　提出模型

本文提出了一种基于优先值的邻居图提取链接预测算法(Priority-based Neighbor Subgraph Extraction method for Link prediction，PNSEL)，后文简称 PNSEL。与 SEAL[14]不同，PNSEL 能更有针对性地提取子图，并且根据提取子图提取时的节点重要性进行编号，从而在目标节点对周围提取出有足够表达力的封闭子图，然后对封闭子图使用图分类算法，预测中心节点对之间是否存在边相连，如图 2 所示。PNSEL 主要包括三个步骤：1)对全图的边进行筛选保留重要性高的边；2)对训练集中的每个节点对提取一个封闭子图并对其中的节点进行节点编号；3)在每一个小图上使用图分类算法进行 0-1 预测。

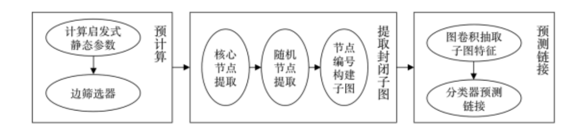


图 2　模型框架图

Fig. 2　General architecture of the proposed model

2.1　问题定义

链接预测任务的目标是，根据已知的图结构数据，预测图中可能存在或者即将出现的其他边。具体的数学定义如下：输入的图数据为 $G=(\mathcal{V},\mathcal{E})$，其中，$\mathcal{V}$ 表示所有的节点集合，$\mathcal{E}$ 表示输入的已知边的集合，其中 $\varepsilon_{i,j}\in\mathcal{E}$ 当且仅当 $v_i,v_j\in\mathcal{V}$ 且 v_i 与 v_j 在输入图数据中之间存在一条边相连。测试集 $\mathcal{E}_{test}$ 是由节点对 (v_i,v_j) 组成的集合，满足 $v_i,v_j\in\mathcal{V}$ 且 $\varepsilon_{i,j}\notin\mathcal{E}$。链接预测任务要解决的问题就是，通过在 G 上的建模和学习，对测试集 $\mathcal{E}_{test}$ 里的节点对之间是否存在连边进行预测。

2.2　边筛选器

边筛选器是对图中的边进行筛选的模块。在非常稠密的图中，总边数的数量级非常大，会导致封闭子图提取步骤的计算量非常大，同时也会使封闭子图占用很大的存储空间。本文可以通过设置边筛选器模块解决这个问题，边筛选器模块可以过滤掉训练集中的一些重要程度不高的边，保留比较核心的边，在保持核心拓扑结构不变的情况下减小计算量，提高算法的效率。具体做法如下：

对于任意的 $\varepsilon_{i,j}\notin\mathcal{E}$，本文计算它的两个端点 v_i,v_j 之间的 Jaccard 系数作为这个边的优先级，即

$$S(v_i,v_j)=\frac{|\Gamma(v_i)\cap\Gamma(v_j)|}{|\Gamma(v_i)\cup\Gamma(v_j)|},\qquad(1)$$

其中，$\Gamma(v)$ 表示节点 v 的一阶邻居节点集合。Jaccard 系数越高，说明两节点之间的关联紧密程度越大，这个边存在的重要性就越高。本文对所有边的 Jaccard 系数进行排序，并保留 $\lfloor k\times|\mathcal{E}|\rfloor$ 条边作为训练集中输入的邻接矩阵，其中 $k\in(0,1]$ 表示保留边的百分比，$\lfloor x\rfloor$ 表不超过 x 的最大整数，新的边集合记为 $\mathcal{E}'$。当 k 取 1 时，本文保留所有的原始边，不进行边筛选。

2.3　封闭子图提取

在本节中，本文提出了一种新的封闭子图提取方法，主要步骤如图 3 所示。这种方法不仅能够选中在目标节点对周围的影响力和重要性高的节点，而且能够保留一定的随机性。随着跳数的扩散，被选入封闭子图的可能性将被随机地分配到目标节点对附近的其他节点上。对于一个给定的目标节点对 (v_i,v_j)，本文先从节点层面出发，在目标节点对周围选择恰当的节点集合 $\mathcal{V}_{i,j}$，从而得到封闭子图的边集合 $\mathcal{E}_{i,j}=\{\varepsilon_{x,y}\,|\,v_x,v_y\in\mathcal{V}_{i,j}$ 且 $\varepsilon_{x,y}\in\mathcal{E}'\}$，即端点均为 $\mathcal{V}_{i,j}$ 中的节点且出现在过滤完的全图边的集合 $\mathcal{E}'$ 中的所有边构成的集合，最终提取的封闭子图就是 $G_{i,j}=(\mathcal{V}_{i,j},\mathcal{E}_{i,j})$。

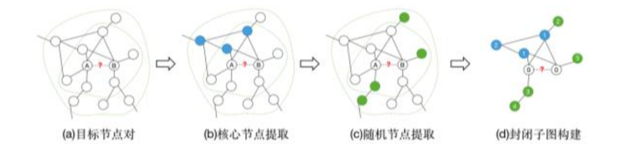


图 3　封闭子图提取步骤

Fig. 3　Extraction steps of closing subgraph

在节点集合的提取过程中，为了使提取的子图兼具影响力和随机性，本文将提取的节点集合分成两个部分：核心节点集合 $\mathcal{V}_{i,j}^{\mathrm{prior}}$ 和随机节点集合 $\mathcal{V}_{i,j}^{\mathrm{rand}}$。他们之间满足这样的关系，$\mathcal{V}_{i,j}=\mathcal{V}_{i,j}^{\mathrm{prior}}\cup\mathcal{V}_{i,j}^{\mathrm{rand}},\mathcal{V}_{i,j}^{\mathrm{prior}}\cap\mathcal{V}_{i,j}^{\mathrm{rand}}=\varnothing$。本文使用超参数 α 来决定封闭子图节点集合的随机性和影响力排序的重要性大小占比，即本文使用核心节点提取方法提取 $\lfloor\alpha\times|\mathcal{V}_{i,j}|\rfloor$ 数量的点，使用随机节点提取方法提取 $\lfloor(1-\alpha)\times|\mathcal{V}_{i,j}|\rfloor$ 数量的点，其中：

$$\alpha=\frac{|\mathcal{V}_{i,j}^{\mathrm{prior}}|}{|\mathcal{V}_{i,j}|}=\frac{|\mathcal{V}_{i,j}^{\mathrm{prior}}|}{|\mathcal{V}_{i,j}^{\mathrm{rand}}|+|\mathcal{V}_{i,j}^{\mathrm{prior}}|}\in[0,1]\qquad(2)$$

2.3.1 核心节点提取

核心节点提取部分旨在提取出相对于目标节点对和全图都具有高影响力的重要节点。在实际操作中，本文使用全局 PageRank 和个性化的 PageRank[17]来表示节点的全局影响力和相对于目标节点对的局部影响力。本文用 $pr_i,v_i\in\mathcal{V}$ 表示全局 PageRank，用 ppr_x^y 表示以节点 v_y 为出发点计算出来的节点 v_x 的个性化的 PageRank，其中 $v_x,v_y\in\mathcal{V}$。那么对于一个固定的目标节点对 (v_i,v_j)，他们的周围节点 v_x 的全局影响力就用 pr_x 表示；v_x 的局部相对影响力用分别以两个目标节点为核心节点计算出来的个性化的 PageRank 的最大值来计算，也就是说节点 v_x 相对于目标节点对 (v_i,v_j) 的局部相对影响力大小为 $\max(ppr_x^i,ppr_x^j)$。同时，本文用超参数 β 来控制局部影响力在核心节点排序评分中的重要性大小，也就是说本文最

后的周围节点优先值计算方法如下：

$$p_x^{i,j} = \beta \times \max(ppr_x^i, ppr_x^j) + (1-\beta) \times pr_x, v_x \in \mathcal{V} \tag{3}$$

然后本文可以通过排序所有节点的优先值得到优先值最高的 $n_1 = [\alpha \times \text{fix_node_num}]$ 个节点，来得到目标节点对的核心节点集合。

2.3.2 随机节点提取

随机节点提取部分旨在随机提取出目标节点对周围的邻居节点(包括直接相邻和间接相邻)。在随机节点提取部分，所提方法采用了类似最小哈希算法(MinHash)[24]的思想，最小哈希算法是利用低维编码的方式快速近似计算两个集合的 Jaccard 相似性的算法。在这个模块中，所提算法将每个筛选出来的节点集合视为一个编码，分别编码目标节点对 (v_i, v_j) 的 p 跳邻居：$\mathcal{N}_i^p, \mathcal{N}_j^p$，其中 $p = 1,2,3\ldots\text{num_hops}$，这样所筛选出来的节点就能很好的代表两个中心节点的邻居特征。

具体来说，随机节点提取模块提取节点的总数量为

$$n_2 = [(1-\alpha) \times \text{fix_node_num}], \tag{4}$$

本文先生成 n_2 次相互独立的全图节点随机排列的哈希函数

$$perm^k : \mathcal{V} \to \mathbb{N} \quad , k = 1,2\cdots n_2 \tag{5}$$

A 函数的输出是 0 到(节点总数-1)上的正整数，输入是图中的某一个节点。同时本文构建一个固定序号哈希函数 $h: \mathbb{N} \to \mathcal{V}$，即每个序号唯一的对应图上的某一个节点。为了均匀的分配提取的随机节点，本文的算法会在每一跳的邻居上采样

$$\text{node_per_hop} = [\frac{n_2}{\text{num_hops}}] \tag{6}$$

个节点，其中 $[\text{node_per_hop}/2]$ 个节点用来编码 v_i，即对于每一个 $p = 1,2,3\ldots\text{num_hops}$，本文使用

$$h_{\min}(\mathcal{N}^p(v_i)) = h(\min_{u \in \mathcal{N}^p(v_i)} perm^k(u)) \tag{7}$$

计算 $[\text{node_per_hop}/2]$ 次，得到选取的节点集合，其中 $\mathcal{N}^p(v_i)$ 表示节点 v_i 的第 p 跳邻居。同样，本文用剩下的 $[\text{node_per_hop}/2]$ 个节点来编码 v_j，即采样函数为

$$h_{min}(\mathcal{N}^p(v_j)) = h(min_{u \in \mathcal{N}^p(v_j)} perm(u)) \tag{8}$$

最后，将这些选中的节点加入随机节点提取集合 $\mathcal{V}_{i,j}^{\text{rand}}$。

2.3.3 节点编号

节点编号部分的任务是，在已经提取好的封闭子图里，给每一个节点按照重要性赋予一个节点编号。为了在有节点特征和无节点特征的图链接预测任务中都进行子图分类训练，并且统一地在不同的封闭子图中学到局部特征结构来预测核心节点之间是否存在边相连，本文需要使用相同的规则给子图中的节点进行编号。在所有的子图进入图卷积神经网络中进行图分类训练之前，本文给相同编号的节点赋予相同的节点特征。节点编号在连接预测中具有非常重要的意义。Zhang 等人[25]最近提出了一种节点编号理论，该理论提出，链接预测任务本质上是基于点集来提取信息特征进行训练和预测的任务。如果本文仅仅关注节点本身的拓扑结构特征，那么就会陷入对称性的陷阱当中。

该理论还定义了一种编号技巧(Labeling Trick)，并证明了在使用图卷积神经网络来训练节点特征的情况下，结合编号技巧来提取点集特征的方法是一种最具表达力的点集结构特征提取方法。

编号技巧的定义如下：给定 $(S, \mathbf{A})$ 作为节点集合和节点-连边特征矩阵，如果一个编号向量 $\boldsymbol{L}^{(S)} \in \mathbb{R}^{n \times n \times d}$ 满足以下条件就可以称为一个编号技巧：对于任意的 $S, \mathbf{A}, S', \mathbf{A}', \pi \in \Pi_n$，均有

a) 目标节点标识性。

$$\boldsymbol{L}^{(S)} = \pi(\boldsymbol{L}^{(S')}) \Rightarrow S = \pi(S') \tag{9}$$

b) 排列变换相等性。

$$S = \pi(S'), \mathbf{A} = \pi(\mathbf{A}') \Rightarrow \boldsymbol{L}^{(S)} = \pi(\boldsymbol{L}^{(S')}) \tag{10}$$

其中，π 是一个排列变换，Π_n 是 n 个元素的所有可能的排列组合。

在跳数更小节点区别性更高的情况下，本文提出了一种新的编号方法，即用核心节点提取模块中计算的目标节点周围节点优先值排序来作为编号：最重要的节点即两个目标节点对，编号为 1，剩下的其他节点按照优先值降序依次编号为 3 至 $n = \text{fix_node_num}$，而其他未被选中的所有节点均编号为 0。下面本文来证明这种编号方法是一种编号技巧：

a) 如果存在 $\boldsymbol{L}^{(S)} = \pi(\boldsymbol{L}^{(S')})$，即 S' 经过变换过的节点编号与 S 完全相同，由于本文除了目标两节点对的编号为 1，其他节点的编号均为一点一个编号，所以肯定可以找一种映射方式 π 使得 S 中的每个节点一一对应到 S' 中编号相同节点。

b) 如果 $S = \pi(S'), A = \pi(A')$，即图 $(S, \mathbf{A})$ 与图 $(S', \mathbf{A}')$ 是同构图，那么以节点 $v_{i'} \in S'$ 为出发点计算的个性化 PageRank 与以节点 $v_i = \pi(v_{i'})$ 为出发点计算的个性化 PageRank 向量必然完全相等，因而由个性化 PageRank 排序得到的节点编号必然也相等。

所以，本文证明了所提的编号方法能够与图卷积神经网络结合，构造出一种最具表达力的点集结构特征提取方法。

2.4 图分类

最后，本文使用图卷积神经网络来对每个构建好的封闭子图进行 0-1 图分类预测。预测为 0 表示目标节点对之间不存在边，预测为 1 表示目标节点对之间存在边。

本文先通过图卷积神经网络提取子图特征

$$g_{i,j} = \text{GNN}(\mathcal{G}_{i,j}) = \text{GNN}((\mathcal{V}_{i,j}, \mathcal{E}_{i,j})), \tag{11}$$

其中，$\text{GNN}(\cdot)$ 表示某一种图卷积神经网络函数。

这里本文主要使用的是 GraphSAGE[11]模型。具体来说，模型先初始化节点特征为图数据的节点原生特征

$$\mathbf{Z}^{(0)} = X, \tag{12}$$

然后通过聚合邻居节点的特征，来逐层更新节点特征

$$\begin{aligned} \mathbf{z}_{\mathcal{N}(v)}^t &= \text{AGGREGATE}_t(\{\mathbf{z}_u^{t-1}, \forall u \in \mathcal{N}(v)\}) \\ \mathbf{z}_v^t &= \sigma(\boldsymbol{W}^t \cdot \text{CONCAT}(\mathbf{z}_v^{t-1}, \mathbf{z}_{\mathcal{N}(v)}^t)) \end{aligned} \tag{13}$$

其中，$t \in \{1,2,\cdots,h-1\}$，本文将两个目标节点 v_i, v_j 的节点特征做哈达玛积(Hadamard product)得到子图的图特征向量

$$\begin{aligned} g_{i,j} &= z_i \odot z_j \\ g_{i,j}[x] &= z_i[x] \times z_j[x] \end{aligned} \tag{14}$$

然后本文将子图特征通过多层感知机，得到链接预测值

$$\begin{aligned} g_{i,j}^{(0)} &= g_{i,j} \\ g_{i,j}^{(k)} &= \text{ReLU}(W_k g_{i,j}^{(k-1)} + b_k), k \in \{1,2,\cdots,K-1\} \\ y_{i,j} &= \sigma(W_K g_{i,j}^{(K-1)} + b_K) \end{aligned} \tag{15}$$

其中，$\sigma(\cdot)$ 是 sigmoid 函数，$y_{i,j}$ 即为本文对 v_i, v_j 的连边情况的预测。

3 实验及分析

本文在有节点特征的数据集和不含节点特征的数据集上分别进行了实验，并与几个基线模型进行了对比实验。下面从数据集、基线模型、评估指标、与基线模型的对比和模型分析讨论等方面对实验和模型进行描述。

3.1 含节点特征数据集

Chameleon，Squirrel 数据集来自维基百科数据集[26]，在这两个数据集中每个节点代表一个网页，而每一条边代表两个不同网页之间的超链接，节点特征则表示网页中存在的特定的代表性的名词含量。

Actor 数据集[27]取自一个“电影-导演-演员-作家”网络，数据集中的每一个节点表示一个演员，如果两个演员在同一个维基百科网页上同时出现过，那么他们会存在一条连边，节点特征反映了该演员的维基百科介绍页面上的一些关键词情况。

Cornell，Texas，Wisconsin 数据集[27]是卡耐基梅隆大学收集整理的不同大学计算机系的校园网页数据集。每个数据集来自一个大学，数据集内的每个节点表示一个网站，网页分为学生、项目、课程、员工和教师这五个类别，节点之间的连边表示网页之间的超链接，节点特征也是网页上出现的关键词信息。

PubMed，Cora，CiteSeer 数据集是非常经典的不同领域的论文引用网络数据集[28,29]。在这三个数据集中，每个节点表示一篇论文，节点之间的连边表示论文之间的互相引用关系，节点特征表示论文的代表词描述信息。

这 9 个含节点特征数据集来自不同的背景，也具有不同的大小规模和平均节点度数，能够很好的综合反映 PNSEL 在不同结构的数据集上的性能，具体的数据规模描述如表 1 所示。

表 1　有节点特征数据集的统计信息

Tab. 1　Statistics of datasets with node features

Dataset	Nodes	Edges	Avg. Degree	Features
Chameleon	2277	36101	15.85	2325
Squirrel	5201	217073	41.74	2089
Actor	7600	33544	4.41	931
Cornell	183	295	1.61	1703
Texas	183	309	1.69	1703
Wisconsin	251	499	1.99	1703
PubMed	19717	44338	2.25	500
Cora	2708	5429	2.00	1433
CiteSeer	3312	4732	1.43	3703

3.2　无节点特征数据集

本文使用了 8 个无节点特征数据集，分别是美国航线数据集 USAir，网络科学研究人员的合作关系网络 NS[30]，美国政治博客网络 PB[31]，蛋白质相互作用网络 Yeast[32]，秀丽隐杆线虫的生物神经网络 C.elegans[33]，美国西部电网分布结构 Power[33]，路由器构建的互联网络图 Router[34]，大肠杆菌中代谢物的成对反映网络 E.coli[35]。他们具有不同的背景、数据规模、平均度数和聚类系数，具体数值分布如表 2 所示。

表 2　无节点特征数据集的统计信息

Tab. 2　Statistics of datasets without node features

Dataset	Nodes	Edges	Avg. Degree
USAir	332	2126	6.40
NS	1589	2742	1.73
PB	1222	16714	13.68
Yeast	2375	11693	4.92
C.ele	297	2148	7.23
Power	4941	6594	1.33
Router	5022	6258	1.25
E.coli	1805	15660	8.68

所有数据集均随机选取原图中 80%的边作为训练集里的可见边，10%的边作为测试集里面的正样本，并随机选取等数量的不存在边相连的节点对作为测试集里面的负样本，剩下的 10%的边作为验证集里的正样本，也同时独立抽取等数量的训练集里不存在边相连的节点对作为验证集里的负样本。

3.3　基线模型

由于所提方法是建立在“提取子图+图分类”框架下的预测算法，目前采用这个框架的算法只有 SEAL 模型，所以 SEAL 是主要的对比对象。近些年来也涌现了一下不同思路的链接预测方法，考虑到算法的角度不同，这里不做对比。本文选择的一些有代表性的模型有：

a) SEAL[14]。该模型使用目标节点对周围的 k 跳邻居小图，通过图分类进行链接预测，是建立在“提取子图+图分类”框架下的目前表现最佳的算法。

b) Node2vec[6]。该模型是一种非常有效的无监督的图嵌入方法，通过随机游走序列学习每个节点的图嵌入表达。训练完成后将目标两节点特征的哈达玛积通过线性层和激活层后进行链接预测。

c) MLP。多层感知机模型，可以使用在含有节点特征的图数据中。模型读入节点的原始特征，将两节点的原始特征的哈达玛积通过深度神经网络后进行 0-1 预测。

d) GraphSAGE[9]。该模型是一种结合邻居采样和动态更新节点特征的图卷积神经网络模型。模型先在全图进行卷积操作更新所有节点的特征，再取出目标节点对的特征向量进行建模预测链接是否存在。

3.4　评估指标

链接预测任务是一种二分类任务，测试集由未知连边情况的节点对组成，其中 50%的节点对在原数据集上存在边相连，但是在训练数据中连边被删去不可见；另外 50%的节点对是随机采样取出的在原图中本来就没有连边的节点对，所以正负样本比例为 1:1。

本文采用 AUC、F1-score、precision 和 recall 作为评价指标，综合的评价预测的准确性。具体计算方法为

$$AUC=\frac{\sum_{i\in \text{PositiveClass}} \text{rank}_i-\frac{M\times(M+1)}{2}}{M\times N} \tag{16}$$

其中，rank_i 表示序号为 i 的样本的预测概率在所有样本从小到大排序后的排序序号，M 、N 表示是正样本和负样本的个数，$PositiveClass$ 表示正样本的序号集合。

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}},\quad \text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{17}$$

其中，precision 即精准率，表示分类器判定的正例中的正样本比例，recall 即召回率，表示正样本中被分类器判定为正例的比例。TP 表示预测为正例的正样本数量，FP 表示预测为正例的负样本数量，FN 表示预测为负例的正样本数量。

$$\text{F1}-\text{score}=2\times\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{18}$$

3.5　参数设置

本文在 9 个含节点特征数据集和 8 个不含节点特征的数据集上对上述的基线模型进行了实验。对于 Node2vec 模型，本文采用的随机游走步长为 10，窗口长度为 5，训练的节点特征维数为 128 维，然后本文使用相同的训练集验证集测试集划分，用 Node2vec 得到的节点特征为输入训练 MLP 分类模型来做链接预测。本文在有节点特征的数据集上使用 MLP 方法作为一个基线模型，使用节点原生特征作为 MLP 模型的输入特征向量，MLP 的层数设置为 3 层，隐藏层的向量维数为 256 维。对于 GraphSAGE 模型，在有节点特征的数据集上，本文采用了两种训练方式。一种是初始化节点特征向量为节点原生特征，固定输入的节点特征向量，这种模型记为 GraphSAGE_1；第二种是随机初始化节点的特征向量，然后将特征向量当成参数进行训练，这种模型记为 GraphSAGE_2。在不含节点特征的数据集上，本文只采用了随机初始化节点特征向量参数，并训练特征向量参数的方式，模型记作 GraphSAGE。GraphSAGE 的卷积层数设置为 2 层，隐藏层的向量维数同样设置为 256 维。对于 SEAL 模型，在原文章中已有实验的数据集，本文采用与 SEAL 论文中相同的实验设置。在原文章中没有的数据集，本文使用与原文章中相似数据集类似的参数实验，并对主要的超参实验了主要可能的取值，选择最佳结果作为实验最终结果。对于本文提出的算法 PNSEL，本文同样使用了 256 维的隐藏层维数，在计算周围节点优先值时，局部影响力占比超参数 β 本文在

[0, 0.3, 0.5, 0.7, 1]这几个数值中进行了实验。在分配核心节点提取比例时，核心节点占比超参数 α 本文在[0, 0.3, 0.5, 0.8, 1]。本文模型使用的编号方式为双半径编号法和优先值排序编号法，使用的图分类模型为 DGCNN[36]和 GraphSAGE，选择最佳结果作为实验的最终结果。PNSEL 与 SEAL 均将 batch size 数量设置为 32。上述模型均使用各数据集上最佳的学习率，训练的 epoch 数均为 100，并进行独立实验 10 次，计算正确率的均值和方差。

3.6 与基线模型的对比

本文分别在有节点特征和无节点特征两个情况下分析本文模型的性能。表 3~8 是在有节点特征的 9 个数据集上的 AUC、F1-score、precision 和 recall 的实验结果(precision 和 recall 的结果为 10 次独立实验的均值)。

表 3 与基线模型在有特征数据集(a)上的比较(AUC)

Tab. 3 Comparasion with baselines on datasets (a) with node feature(AUC)

method	Actor	Chameleon	Citeseer	Cora
PNSEL	84.86 ± 0.24	99.78 ± 0.01	90.07 ± 0.08	91.59 ± 0.24
SEAL	75.28 ± 0.56	99.60 ± 0.06	90.53 ± 0.84	90.67 ± 0.02
Node2vec	78.52 ± 0.69	98.28 ± 0.02	78.34 ± 0.35	86.08 ± 0.55
MLP	53.03 ± 0.23	97.14 ± 0.05	91.51 ± 0.12	82.58 ± 0.76
GraphSAGE1	82.00 ± 0.16	99.66 ± 0.02	92.69 ± 0.37	93.88 ± 0.47
GraphSAGE2	80.72 ± 0.61	99.27 ± 0.02	72.61 ± 2.26	77.97 ± 1.37

表 4 与基线模型在有特征数据集(a)上的比较 (F1-score)

Tab. 4 Comparasion with baselines on datasets(a) with node feature(F1-score)

method	Actor	Chameleon	Citeseer	Cora
PNSEL	76.24 ± 0.62	98.34 ± 0.06	82.80 ± 0.44	82.81 ± 0.91
SEAL	68.25 ± 0.65	97.86 ± 0.28	80.15 ± 0.05	82.19 ± 0.77
Node2vec	66.37 ± 1.04	94.37 ± 0.13	69.80 ± 2.07	73.62 ± 6.24
MLP	53.02 ± 0.76	92.06 ± 0.17	83.97 ± 0.83	73.23 ± 2.70
GraphSAGE1	38.11 ± 2.02	97.17 ± 0.08	53.27 ± 2.36	65.31 ± 0.94
GraphSAGE2	52.98 ± 2.12	96.46 ± 0.05	40.33 ± 1.62	43.86 ± 2.00

表 5 与基线模型在有特征数据集(a)上的比较 (Precision, Recall)

Tab. 5 Comparasion with baselines on datasets(a) with node feature

method	Actor		Chameleon		Citeseer		Cora	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
PNSEL	79.38	73.45	98.30	**98.38**	81.68	83.96	87.33	78.80
SEAL	69.84	66.91	97.51	97.80	90.90	71.72	91.55	74.57
Node2vec	52.31	**91.06**	97.27	91.64	68.87	76.44	68.71	**86.07**
MLP	52.46	53.62	89.40	94.91	81.62	**86.90**	80.99	67.51
$GraphSAGE_1$	**95.88**	23.79	**98.58**	95.80	**99.40**	36.41	**98.85**	48.77
$GraphSAGE_2$	92.31	37.18	97.94	95.02	92.77	25.79	94.92	28.53

表 6 与基线模型在有特征数据集(b)上的比较(AUC)

Tab. 6 Comparasion with baselines on datasets(b) with node feature(AUC)

method	Cornell	PubMed	Squirrel	Texas	Wisconsin
PNSEL	88.51 ± 2.10	97.70 ± 0.02	99.76 ± 0.01	84.45 ± 2.57	81.63 ± 6.18
SEAL	82.24 ± 2.69	97.37 ± 0.37	99.64 ± 0.32	81.25 ± 3.85	71.01 ± 1.97
Node2vec	63.81 ± 6.81	80.14 ± 0.65	98.87 ± 0.01	62.42 ±10.09	60.76 ± 5.57
MLP	74.12 ± 0.90	91.97 ± 0.20	95.77 ± 0.05	76.17 ± 7.50	82.23 ± 5.95
GraphSAGE1	76.59 ± 3.41	92.99 ± 0.15	99.29 ± 0.03	79.95 ± 1.26	79.62 ± 1.05
GraphSAGE2	76.18 ± 7.21	93.09 ± 0.06	99.46 ± 0.01	78.84 ± 3.83	76.56 ± 4.63

如表中所示，对于有节点特征的数据，PNSEL 在社交网络、论文引用、生物关联等网络关系数据中与其他基线模型相比，都表现出了最佳的平均 AUC 和 F1-score，同时，precision 和 recall 的表现也非常均衡， precision 和 recall 综合来看的平均情况最佳，这说明本文的模型具有很好的性能和优秀的适应性。

Node2vec 模型在部分数据上也有比较优良的表现，说明无监督的图嵌入方法也能在一定程度上提取出链接预测需要用到的拓扑信息，但是在其他的大部分数据上则无法保持很好的效果。MLP 模型在 chameleon、Wisconsin、CiteSeer 上表现也不错，其中 Wisconsin 的预测 AUC 比其他基线模型都高，说明对于某些含节点特征的图网络结构而言，节点的原生特征对于链接预测起到了首要的作用。GraphSAGE 模型在使用和不使用原生特征的情况下，总的来说模型表现差异不大，在大部分数据集上均能有不错的效果，其中在 CiteSeer 和 Cora 上的 AUC 完全超过其他基线模型但是 F1-score 却比较低，说明图神经网络模型对于链接预测来说的表达力很强，但是存在正负例的预测准确性不均衡的问题。SEAL 模型综合了以上模型的优点，是所有数据集上平均表现第二好的模型，说明“子图提取+图分类”的框架在链接预测问题上具有非常好的效果，但是仍有一定的提升空间。而 PNSEL 在大部分的数据集上均表现出了显著高于 SEAL 的 AUC 和 F1-score，说明本文的改进是非常有效且合理的。

表 7 与基线模型在有特征数据集(b)上的比较(F1-score)

Tab. 7 Comparasion with baselines on datasets(b) with node feature(F1-score)

method	Cornell	PubMed	Squirrel	Texas	Wisconsin
PNSEL	78.77 ± 3.06	92.66 ± 0.04	97.87 ± 0.10	69.62 ± 4.07	71.80 ± 5.40
SEAL	72.80 ± 5.31	91.30 ± 0.78	97.21 ± 0.52	49.70 ± 8.40	42.51 ± 36.93
Node2vec	55.56 ± 24.85	68.55 ± 0.95	95.42 ± 0.09	12.73 ± 28.46	66.67 ± 0.00
MLP	67.42 ± 4.59	84.93 ± 0.36	88.29 ± 0.07	69.00 ± 4.47	71.75 ± 10.35
GraphSAGE1	29.75 ± 29.63	76.12 ± 1.44	95.99 ± 0.29	27.15 ± 13.71	41.25 ± 5.48
GraphSAGE2	17.74 ± 5.74	62.52 ± 0.89	96.71 ± 0.08	45.43 ± 26.35	39.63 ± 9.58

表 8 与基线模型在有特征数据集(b)上的比较(Precision, Recall)

Tab. 8 Comparasion with baselines on datasets(b) with node feature

method	Cornell		PubMed		Squirrel		Texas		Wisconsin	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
PNSEL	83.85	74.71	90.64	94.77	98.35	97.4	92.38	56.25	77.23	70.00
SEAL	86.75	62.75	93.21	89.47	98.02	96.41	100.0	33.33	44.85	40.58
Node2vec	60.00	81.18	98.80	52.50	97.89	93.06	10.00	17.50	50.00	100.0
MLP	84.40	56.47	82.58	87.48	90.31	86.35	53.71	97.50	92.10	60.00
GraphSAGE1	60.00	21.57	97.99	62.26	98.57	93.55	93.33	16.67	100.0	26.09
GraphSAGE2	100.0	9.80	97.49	46.02	98.62	94.86	96.97	33.33	85.00	26.09

表 9~11 是在不含节点特征的 8 个数据集上的 AUC、F1-score、precision 和 recall 的实验结果(precision 和 recall 的结果为 10 次独立实验的均值)。对于不含节点特征的数据集，Node2vec 可以比较好的学到数据中的结构信息，与 GraphSAGE 的表现不相上下，但与表现最佳的模型仍有一定的显著差距。这说明在不存在节点特征的情况下，这两种模型在不同背景不同特质的图数据中不能稳定预测链接是否存在。SEAL 明显的表现由于另外两个模型，同时 PNSEL 也相比 SEAL 有显著的性能提升，说明所提模型不仅在对节点特征利用上有更好的性能，而且在不存在节点特征的情况下，PNSEL 也能更有效地利用子图信息作出预测。

表 9 与基线模型在无特征数据集上的比较(AUC)

Tab. 9 Compare with baselines on datasets without node feature(AUC)

Dataset	PNSEL	SEAL	Node2vec	GraphSAGE
C.ele	**90.33 ± 0.21**	82.44 ± 0.82	74.12 ± 0.37	86.75 ± 0.72
E.coli	**97.74 ± 0.03**	95.33 ± 0.12	94.50 ± 0.06	94.40 ± 0.07
NS	**98.17 ± 0.04**	91.18 ± 1.37	94.04 ± 0.08	81.11 ± 2.07
PB	**94.94 ± 0.02**	92.71 ± 0.05	89.53 ± 0.32	94.37 ± 0.04
Power	**92.28 ± 0.10**	72.21 ± 1.38	80.16 ± 0.52	64.97 ± 3.24
Router	**94.64 ± 0.34**	81.86 ± 0.60	76.08 ± 2.35	77.68 ± 2.33
USAir	**96.55 ± 0.11**	94.05 ± 0.65	84.57 ± 0.64	94.62 ± 0.73
Yeast	**97.87 ± 0.05**	91.71 ± 0.13	94.07 ± 0.24	93.85 ± 0.38

3.7 模型分析和讨论

本节通过控制改变单一维度的参数进行实验来说明本文

提出的算法中的关键部分对模型的效果及作用。

表 10　与基线模型在无特征数据集上的比较(F1-score)

Tab. 10　Comparasion with baselines on datasets without node feature(F1-score)

Dataset	PNSEL	SEAL	Node2vec	GraphSAGE
C.ele	**82.95 ± 0.91**	73.96 ± 1.16	61.56 ± 2.45	71.32 ± 5.34
E.coli	**92.95 ± 0.23**	89.37 ± 0.15	87.49 ± 1.23	88.22 ± 0.14
NS	**94.03 ± 0.12**	83.21 ± 2.05	91.06 ± 0.81	70.98 ± 0.30
PB	**88.65 ± 0.30**	84.85 ± 0.54	75.22 ± 7.90	87.78 ± 0.41
Power	**84.14 ± 0.48**	54.59 ± 0.70	69.00 ± 2.44	15.27 ± 2.24
Router	**86.04 ± 0.49**	73.80 ± 1.18	66.34 ± 0.33	37.17 ± 0.76
USAir	**90.18 ± 0.66**	88.24 ± 0.76	68.93 ± 3.30	84.38 ± 1.76
Yeast	**93.48 ± 0.26**	86.19 ± 0.28	82.00 ± 9.37	85.44 ± 0.07

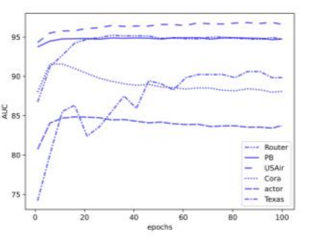
表 11　与基线模型在无特征数据集上的比较(Precision, Recall)

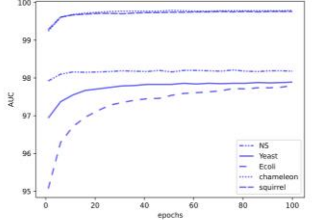
Tab. 11　Comparison with baselines on datasets without node feature

Dataset		PNSEL	SEAL	Node2vec	GraphSAGE
C.ele	Pre	79.47	78.50	76.46	**86.06**
	Rec	**86.76**	69.94	51.59	61.21
E.coli	Pre	95.80	92.58	95.86	**96.66**
	Rec	**90.27**	86.43	80.52	81.15
NS	Pre	**97.27**	88.31	95.34	94.93
	Rec	**91.00**	78.95	87.23	56.69
PB	Pre	87.67	**89.51**	67.04	87.27
	Rec	89.67	80.65	**91.38**	88.31
Power	Pre	80.82	81.68	**91.51**	86.41
	Rec	**87.76**	41.02	55.99	8.40
Router	Pre	86.02	84.06	50.13	**97.07**
	Rec	86.13	65.81	**98.05**	22.99
USAir	Pre	90.12	89.78	53.49	**96.02**
	Rec	90.25	86.79	**97.74**	75.31
Yeast	Pre	95.00	92.97	75.15	**97.76**
	Rec	92.02	80.33	**93.17**	75.88

3.7.1 收敛性分析

AUC 是模型训练时的主要指示性指标，本文通过绘制训练次数与正确率(AUC)变化曲线来观察模型的收敛情况，如图 4 所示。





(a)低正确率数据集收敛曲线　　(b)高正确率数据集收敛曲线

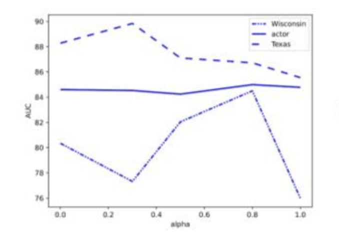
图 4　正确率收敛曲线

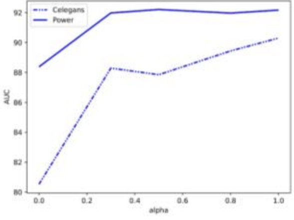
Fig. 4　Convergence of AUC

本文选择了 5 个有特征数据集和 6 个无特征数据集的训练情况进行绘图，分别用虚线和实线表示。由于正确率的量级不同，本文将曲线分开绘制在两张图上。从图上可以看出，在所有的数据集上，虽然正确率的收敛速度和曲线形状有所不同，但是最终正确率都收敛到了某一特定值，说明 PNSEL 具有良好的收敛性。

3.7.2 核心节点占比参数影响

核心节点比重参数 α 表示的是在封闭子图提取时，按节点优先值提取的节点数量占节点总数的比例。本文选取了三个含节点特征数据集和两个不含节点特征的数据集进行实验，在各自最佳参数的其他参数保持不变的情况下，改变 α 进行实验，独立进行 3 次实验取平均结果，如图 5 所示。





(a)有特征数据集　　（b）无特征数据集

图 5　不同的核心节点占比参数的影响

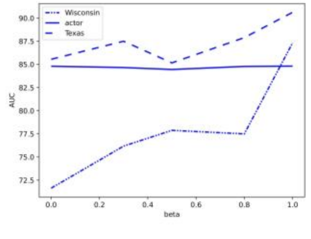
Fig. 5　Influence of alpha

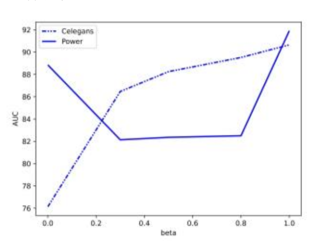
可以看出，核心节点比重参数对链接预测任务的准确性在不同的数据集上均具有明显的影响。其中，在含节点特征的数据集上，不同的数据往往有着不同的最佳 α 选择，过大或者过小均不能达到最佳的预测准确性。这说明这些图结构更适合核心节点抽取与随机节点抽取结合的方式。

而在不含节点特征的数据集上，本文注意到他们的最佳水平往往出现在 α 为 1 时，且随着 α 变大预测准确性大致呈上升趋势，说明在这类不含节点特征的数据集上，完全使用核心节点抽取是最佳的提取封闭子图的方法。

3.7.3 局部影响力占比参数影响

局部影响力占比参数 β 表示的是本文在计算周围节点优先值时，局部影响力特征所占的比重。β 越大，节点的局部影响力在优先值里的比重就越大，节点的全图影响力在优先值里的比重就越小。本文选取了同样三个含节点特征数据集和两个不含节点特征的数据集进行实验，在各自最佳参数的其他参数保持不变的情况下，改变 β 进行实验，独立进行 3 次实验取平均结果，结果如图 6 所示。





(a)有特征数据集　　(b)无特征数据集

图 6　不同的局部影响力占比参数的影响

Fig. 6　Influence of beta

本文可以看到，除了 actor 数据对于 β 不敏感之外，其他数据集的准确性均受 β 的取值影响比较明显。其中，在 Wisconsin、Texas 和 Celegans 上，模型的正确率均随着 β 的增加而增加。这说明对于这些数据而言，考虑节点优先值时仅考虑局部影响力是最佳选择，全局影响力对于局部的节点预测准确性意义不大。而在 Power 数据集上，本文观测到，当 β 取 0 或者 1 时，模型具有最佳表现，说明仅选取全图最重要的节点或者仅选取局部重要的节点都能为链接预测提供足够的有效信息，而两者结合反而会使得筛选变得低效。

4　结束语

本文对链接预测领域目前表现最佳的模型提出了一种改进的方案，基于“子图提取+图分类”的链接预测结构，提出了一种基于优先值的邻居子图提取连接预测算法(PNSEL)。所提算法可以在固定小图规模的情况下，结合局部图结构和全图结构信息，选出对于目标节点对最为重要的周围节点，并保留了一定的随机性以应对差异化的图结构。通过大量在不同背景的真实数据集上的实验，本文将 PNSEL 与具有代表性的几个基线模型进行对比，证明了 PNSEL 相比改进前能显著带来正确率的提高，同时也通过拆解实验证明了主要参数的影响性。

参考文献：

[1] Qi Yanjun, Bar - Joseph Z, Klein - Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction [J]. Proteins: Structure, Function, and Bioinformatics, 2006, 63 (3): 490-500.

[2] Stanfield Z, Coşkun M, Koyutürk M. Drug response prediction as a link prediction problem [J]. Scientific reports, 2017, 7 (1): 1-13.

[3] Adamic L A, Adar E. Friends and neighbors on the web [J]. Social networks, 2003, 25 (3): 211-230.

[4] 王星，王硕，陈吉，侯磊. 联合图注意力和卷积神经网络的链接预测方法 [J]. 山西大学学报 (自然科学版), 2021, 44 (03): 462-470. DOI: 10. 13451/j. sxu. ns. 2020153.

[5] 张玲玲，陈卫静. 合作网络中关键研发者潜在重要合作者链接预测 [J]. 情报探索, 2022 (03): 19-25.

[6] Grover A, Leskovec J. node2vec: Scalable feature learning for networks [C]// Proc of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016: 855-864.

[7] Kipf T N, Welling M. Variational graph auto-encoders [EB/OL]. (2016-11-21) . https://arxiv. org/abs/1611. 07308.

[8] 郝宵荣，王莉，廉涛. 基于节点表示和子图结构的动态网络链接预测 [J]. 模式识别与人工智能, 2021, 34 (02): 117-126. DOI: 10. 16451/j. cnki. issn1003-6059. 202102003.

[9] Hamilton W, Ying Zhitao, Leskovec J. Inductive representation learning on large graphs [C]// Advances in neural information processing systems. 2017: 1025–1035.

[10] Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks [EB/OL]. (2017-10-30) . https://arxiv. org/abs/1710. 10903.

[11] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [EB/OL]. (2016-09-09) . https://arxiv. org/abs/1609. 02907.

[12] Singh A, Huang Qian, Huang S L, *et al.* Edge proposal sets for link prediction [EB/OL]. (2021-06-30) . https://arxiv. org/abs/2106. 15810.

[13] Li Boning, Xia Yingce, Xie Shufang, *et al.* Distance-Enhanced Graph Neural Network for Link Prediction [EB/OL]. (2021) . https://icml-compbio. github. io/2021/papers/WCBICML2021_paper_52. pdf

[14] Zhang Muhan, Chen Yixin. Link prediction based on graph neural networks [C]// Advances in neural information processing systems. 2018: 5165–5175.

[15] Barabási A L, Albert R. Emergence of scaling in random networks [J]. science, 1999, 286 (5439): 509-512.

[16] Zhou Tao, Lyu Linyuan, Zhang Y C. Predicting missing links via local information [J]. The European Physical Journal B, 2009, 71 (4): 623-630.

[17] Page L, Brin S, Motwani R, *et al.* The PageRank citation ranking: Bringing order tothe web, 1999-66 [R]. [S. l.] : Stanford InfoLab, 1999.

[18] Kovács I A, Luck K, Spirohn K, *et al.* Network-based prediction of protein interactions [J]. Nature communications, 2019, 10 (1): 1-8.

[19] Katz L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18 (1): 39-43.

[20] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations [C]// Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 701-710.

[21] Tang Jian, Qu Meng, Wang Mingzhe, *et al.* Line: Large-scale information network embedding [C]// Proc of the 24th International Conference onWorld Wide Web. 2015: 1067-1077.

[22] Xu Keyulu, Hu Weihua, Leskovec J, *et al.* How powerful are graph neural networks? [EB/OL]. (2018-10-01) . https://arxiv. org/abs/1810. 00826.

[23] Leman A A, Weisfeiler B. A reduction of a graph to a canonical form and an algebra arising during this reduction [J]. Nauchno-Technicheskaya Informatsiya, 1968, 2 (9): 12-16.

[24] Broder A Z. On the resemblance and containment of documents [C]// Proc of Compression and Complexity of Sequences 1997. IEEE, 1997: 21-29.

[25] Zhang Muhan, Li Pan, Xia Yinglong, *et al.* Labeling Trick: A Theory of Using Graph Neural Networks for Multi-Node Representation Learning [C]// Advances in Neural Information Processing Systems, 2021.

[26] Rozemberczki B, Allen C, Sarkar R. Multi-scale attributed node embedding [J]. Journal of Complex Networks, 2021, 9 (2): cnab014.

[27] Pei Hongbin, Wei Bingzhe, Chang K C C, *et al.* Geom-gcn: Geometric graph convolutional networks [EB/OL]. (2020-02-13) . https://arxiv. org/abs/2002. 05287.

[28] Sen P, Namata G, Bilgic M, *et al.* Collective classification in network data [J]. AI magazine, 2008, 29 (3): 93-93.

[29] Yang Zhilin, Cohen W, Salakhudinov R. Revisiting semi-supervised learning with graph embeddings [C]// International conference on machine learning. PMLR, 2016: 40-48.

[30] Newman MEJ. Finding community structure in networks using the eigenvectors of matrices [J]. Physical review E, 2006, 74 (3): 036104.

[31] Ackland R. Mapping the US political blogosphere: Are conservative bloggers more prominent? [C]// BlogTalk Downunder 2005 Conference, Sydney, 2005.

[32] Von Mering C, Krause R, Snel B, *et al.* Comparative assessment of large-scale data sets of protein–protein interactions [J]. Nature, 2002, 417 (6887): 399-403.

[33] Watts D J, Strogatz S H. Collective dynamics of‘small-world’networks [J]. nature, 1998, 393 (6684): 440-442.

[34] Spring N, Mahajan R, Wetherall D. Measuring ISP topologies with Rocketfuel [J]. ACM SIGCOMM Computer Communication Review, 2002, 32 (4): 133-145.

[35] Zhang Muhan, Cui Zhicheng, Oyetunde T, *et al.* Recovering metabolic networks using a novel hyperlink prediction method [J]. [EB/OL]. (2016-10-21) . https://arxiv. org/abs/1610. 06941.

[36] Zhang Muhan, Cui Zhicheng, Neumann M, *et al.* An end-to-end deep learning architecture for graph classification [C]// Proc of the 32nd AAAI conference on artificial intelligence. 2018: 4438-4445.